Notes

**Tokenization** 🡪 Breaking down text into words and sentences

**Stopword Removal** 🡪 Filtering common words (e.g. Mary had a little lamb => Mary little lamb)

**N-Grams** 🡪 Identifying commonly occurring groups of words (e.g. New York)

**Word Sense Disambiguation** 🡪 Identifying the context in which the word occurs (e.g. cool water/ cool effects)

**Parts of Speech Tagging** 🡪 (Noun, Verb, Adjective, etc.)

**Stemming** 🡪 Removing ends of words in order to treat them as the same word( **clos**e/ **clos**ed / **clos**ely)

**ALGORITHMS**

Classification 🡪 Naïve Bayes

🡪 Support Vector Machines

Clustering 🡪 K-Means

🡪 Hierarchical Clustering